VQEG

THE VIDEO QUALITY EXPERTS GROUP

RRNR-TV Group
TEST PLAN

Draft Version 1.4

Contact: J. Baïna Tel: +33 (0)3 87 20 75 99
Fax: +33 (0)3 87 76 30 62
E-Mail: jamal.baina@tdf.fr

Editorial History

Version	Date	Nature of the modification
1.0	01/09/2000	Draft version 1, edited by J. Baïna
1.0a	12/14/2000	Initial edit following RR/NR meeting 12-13 December 2000, IRT, Munich.
1.1	03/19/2001	Draft version 1.1, edited by H. R. Myler
1.2	5/10/2001	Draft version 1.2, edited by A.M. Rohaly during VQEG meeting 7-11 May 2001, NTIA, Boulder
1.3	5/25/2001	Draft version 1.3, edited by A.M. Rohaly, incorporating text provided by S. Wolf as agreed upon at Boulder meeting
1.4	26/2/02	Draft version 1.4, prepared at Briarcliff meeting.
1.4a	6/2/02	Replaced Sec. 3.3.2 with text written by Jamal and sent to Reflector

Summary

List of Acronyms

Definitions

List of Acronyms

ANOVA	ANalysis Of VAriance
ASCII	ANSI Standard Code for Information Interchange
CCIR	Comite Consultatif International des Radiocommunications
CODEC	Coder-Decoder
CRC	Communications Research Center (Canada)
DVB-C	Digital Video Broadcasting-Cable
FR	Full Reference
GOP	Group of Pictures
HRC	Hypothetical Reference Circuit
IRT	Institut Rundfunk Technische (Germany)
ITU	International Telecommunications Union
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Motion Pictures Expert Group
NR	No (or Zero) Reference)
NTSC	Nat'l Television Standard Code (60 Hz TV)
PAL	(50 Hz TV)
PS	Program Segment
QAM	Quadrature Amplitude Modulation
QPSK	Quadrature Phase Shift Keying
RR	Reduced Reference
SMPTE	Society of Motion Picture and Television Engineers
SRC	Source Reference Channel or Circuit
SSCQE	Single Stimulus Continuous Quality Evaluation
VQEG	Video Quality Experts Group
VTR	Video Tape Recorder

Definitions

Data collection synchronization

1. Introduction

This document defines the procedure for evaluating the performance of objective video quality models submitted to the Video Quality Experts Group (VQEG) RRNR-TV formed from experts of ITU-T Study Groups 9 and ITU-R Study Group 6. It is based on discussions from the VQEG meeting March 13-17, 2000 in Ottawa, Canada at CRC, the ad hoc RRNR-TV group meeting December 11-15, 2000 in Munich, Germany at IRT and the VQEG meeting May 7-11, 2001 in Boulder, USA at NTIA.

The key goal of this test is to evaluate video quality metrics (VQMs) that emulate single stimulus continuous quality evaluation (SSCQE) with compensation for viewer reaction times (viewer delay + slider performance) and objective amplitude scaling. The evaluation performance tests will be based on the comparison of the SSCQE MOS and the MOSp predicted by models. MOS samples will be delivered every 0.5 second for long sequences.

The goal of VQEG RRNR-TV is to evaluate video quality metrics (VQMs). At the end of this test, VQEG will provide the ITU and other standards bodies a final report (as input to the creation of a recommendation) that contains VQM analysis methods and cross-calibration techniques (i.e., a unified framework for interpretation and utilization of the VQMs) and test results for all submitted VQMs. VQEG expects these bodies to use the results together with their application-specific requirements to write recommendations. Where possible, emphasis should be placed on adopting a common VQM for both RR and NR.

In order to achieve this goal, the purpose of the RRNR-TV Group is to produce a more discriminating test than was accomplished in the VQEG Phase I tests[1]. The quality range of this test will address secondary distribution television. The objective models will be tested using a set of digital video sequences selected by the VQEG RRNR-TV group. The test sequences will be processed through a number of hypothetical reference circuits (HRC's). The quality predictions of the submitted models will be compared with subjective ratings from human viewers of the test sequences as defined by this Test Plan. The set of sequences will cover both 50 Hz and 60 Hz formats. Several bit rates of reference channel are defined for the model, these being zero (No Reference), 10 Kb/s, 56 Kb/s and 256 Kb/s. Model performance will be compared separately with the results from each of the four bit rates, then compared between them.

[1] VQEG Test Plan, Phase I final report <<XXX correct>>

2. Subjective Evaluation Procedure

2.1. The SSCQE method

2.1.1. General description

The single stimulus continuous quality evaluation (SSCQE) method presents a digital video sequence once to the subjective assessment viewer. The video sequences may or may not contain impairments. For this evaluation one of the HRCs will be the Reference sequence (not processed), such that a hidden reference procedure is implemented (see section xxx). Hidden reference implies that the subject is not aware that he/she is evaluating the reference or processed sequence. Subjects evaluate the picture quality in real time using a slider device with a continuous grading scale composed of the adjectives Excellent, Good, Fair, Poor and Bad. This approach is consistent with real-time video broadcasting where a reference sample with no degradation is not available to the viewer explicitly.

2.1.2. Test Design

The test design is a full factor, balanced design to allow analysis of variance (ANOVA). The following presents a brief overview of the test design for each video format (i.e., 525-line, 625-line):

1. 6 Scenes - carefully selected 1-minute segments.
2. 10 HRCs - 1 original, and 9 processed versions. The goal is to obtain uniform distribution across the SSCQE quality scale.

This will produce a total of 60 minutes of SSCQE video. To assure that all the viewers see all the video, each subject will view these 60 minutes of video using two 30-minute sessions, separated by a break.

Multiple randomizations are desired so we will need to edit more than 2 viewing tapes. This randomization should be performed at the clip level (i.e., the ordering of each one minute scene x HRC should be randomized). Preferably, 2 sets of tapes should be used (lets call these the red, and green). Subjects should be randomly assigned to one of the 4 possible orderings (R1-R2, R2-R1, G1-G2, G2-G1). Each lab should have an equal number of subjects at each ordering: perhaps 4 subjects per ordering, for a total of 16 viewers per lab.

The first 6 seconds of each clip should be discarded to allow for stabilization of the viewer's responses. This leaves 54 seconds from each video clip to be considered for data analysis, or 60 clips of 54 seconds each.

2.1.3. Viewing conditions

Viewing conditions should comply with those described in International Telecommunications Union Recommendation ITU-R BT.500-10. An example schematic of a viewing room is shown in Figure 1. Specific viewing conditions for subjective assessments in a laboratory environment are:

- Ratio of luminance of inactive screen to peak luminance: ≤ 0.02
- Ratio of the luminance of the screen, when displaying only black level in a completely dark room, to that corresponding to peak white: ≈ 0.01
- Display brightness and contrast: set up via PLUGE (see Recommendations ITU-R BT.814 and ITU-R BT.815)

- Maximum observation angle relative to the normal[2]: 30^0
- Ratio of luminance of background behind picture monitor to peak luminance of picture: ≈ 0.15
- Chromaticity of background: D_{65}
- Other room illumination: low

The monitor to be used in the subjective assessments is a 19 in. (minimum) professional-grade monitor. For example, a Sony BVM-20F1U or equivalent.

The viewing distance of 4H selected by VQEG falls in the range of 4 to 6 H, i.e. four to six times the height of the picture tube, compliant with Recommendation ITU-R BT.500-10. Soundtrack will not be included.

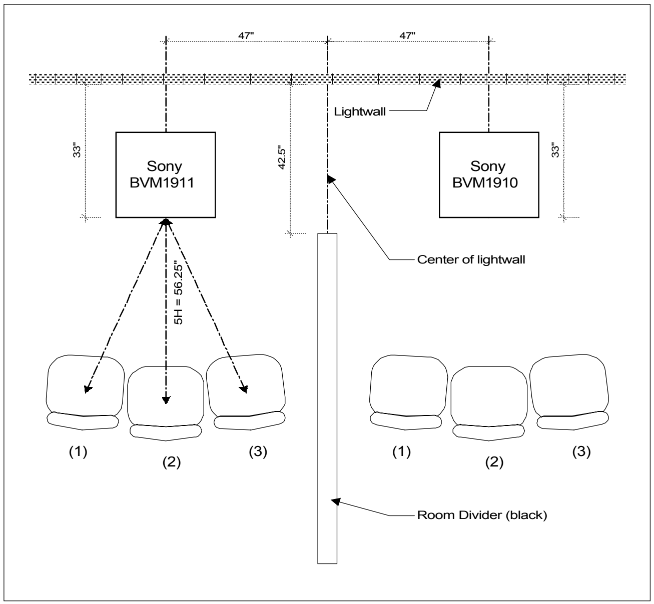

Figure 1. Example of viewing room.

N.B. figure needs to be changed to replace 5H with 4H.

Instructions to viewers for quality tests

The following text should be the instructions given to subjects.

In this test, we ask you to continuously evaluate the video quality of a set of video scenes. The judgment scale shown on the voting device in front of you is a vertical line that is divided into five equal segments. As a guide, the adjectives "excellent", "good", "fair", "poor", and "bad" have been aligned with the five segments of the scale. The quality of the video that you will see may change rapidly and span a range of quality from excellent to bad. During the presentation, you are encouraged to move the indicator along the scale as soon as you notice a change in the quality of the video. The indicator should always be at the point on the scale that currently and accurately corresponds to your judgment of the presentation. You are allowed to move the indicator to any point on the scale. Please do not base your opinion on the content of the scene or the quality of the acting. Take into account the

[2] This number applies to CRT displays, whereas the appropriate numbers for other displays are under study.

different aspects of the video quality and form your opinion based upon your total impression of the video quality.

Possible problems in quality include:

- poor, or inconsistent, reproduction of detail;
- poor reproduction of colors, brightness, or depth;
- poor reproduction of motion;
- imperfections, such as false patterns, blocks, or “snow”.

In judging the overall quality of the presentations, we ask you to use a judgment scale like the sample shown in Figure 2.

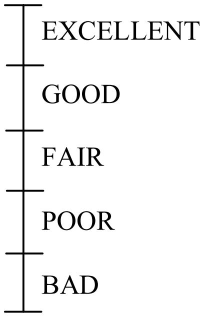

Figure 2. Sample quality scale.

Now we will show a short practice session to familiarize you with the slider operation and the kinds of video impairments that may occur. You will be given an opportunity after the practice session to clarify any questions that you might have. Now please move your slider to the middle position of the quality scale before the practice session begins.

[Run practice session. After the practice session, the test conductor makes sure the subjects understand the instructions and answers any question the subjects might have.]

Before we begin the actual test, please re-position the slider to the middle position of the scale now. We will begin the test in a moment.

[Run the session.]

This completes the test. Thank you for participating.

2.1.4. Viewers

A minimum of 20-25 non-expert viewers should be used. The term non-expert is used in the sense that the occupation of the viewer does not involve television picture quality and they are not experienced assessors. All viewers will be screened prior to participation for the following:

- normal (20/20) visual acuity or corrective glasses (per Snellen test or equivalent)
- normal color vision (per Ishihara test or equivalent)
- sufficient familiarity with language to comprehend instructions and to provide valid responses using semantic judgment terms expressed in that language.

Viable results of at least 16 viewers per lab are required, with viewers equally distributed across sequence randomizations. The subjective labs will agree on a common method of screening the data for validity. Consequently, an additional test is necessary if the number of viewers is reduced to less than 16 per lab as a result of the screening.

2.2. Data format

2.2.1. Results data format

Depending on the facility conducting the evaluations, data entries may vary, however the structure of the resulting data should be consistent among laboratories. An ASCII format data file should be produced with certain header information followed by relevant data. Files should conform to ITU-R Recommendation BT 500-10, Annex 3.

In order to preserve the way in which data is captured, one file will be created with the following information:

Test name: tape number:
Vote type: SSCQE
Lab number:
Viewer number:
Votes number:
Min number:
Max number:

Presentation: Test condition: Program segment:

Time Code	Subject Number 1's opinion	Subject Number 2's opinion	Subject Number 3's opinion
00:00:00:00	…	…	…
00:00:00:12	…	…	…

All these files should have the extension: `.dat` and should be in ASCII format.

2.2.2. Subject data format

The purpose of this file is to contain all information pertaining to individual subjects who participate in the evaluation. The structure of the file should be the following:

Lab Number	Subject Number	Month	Day	Year	Age	Gender*
1	1	07	15	2000	32	1
1	2	07	15	2000	25	2

*Gender where 1=Male, 2=Female

2.2.3. Subjective Data analysis

The subjective test results will be edited to remove the first six seconds of data recorded for each test condition (source/HRC combination). After editing, the validity of the subjective test results will be verified by

1. conducting a repeated measures Analysis of Variance (ANOVA) to examine the main effects of key test variables (source sequence, HRC, etc.),

2. computing means and standard deviations of subjective results from each lab for lab to lab comparisons and

3. computing lab to lab correlations as done for the first VQEG test (ref. VQEG Final Report.).

Once verified, overall means and standard deviations of subjective results will be computed to allow comparison with the outputs of objective models (see section 5).

Data analysis will be conducted over normalized and non-normalized data sets.

3. Sequence processing and data formats

3.1. Sequence processing overview

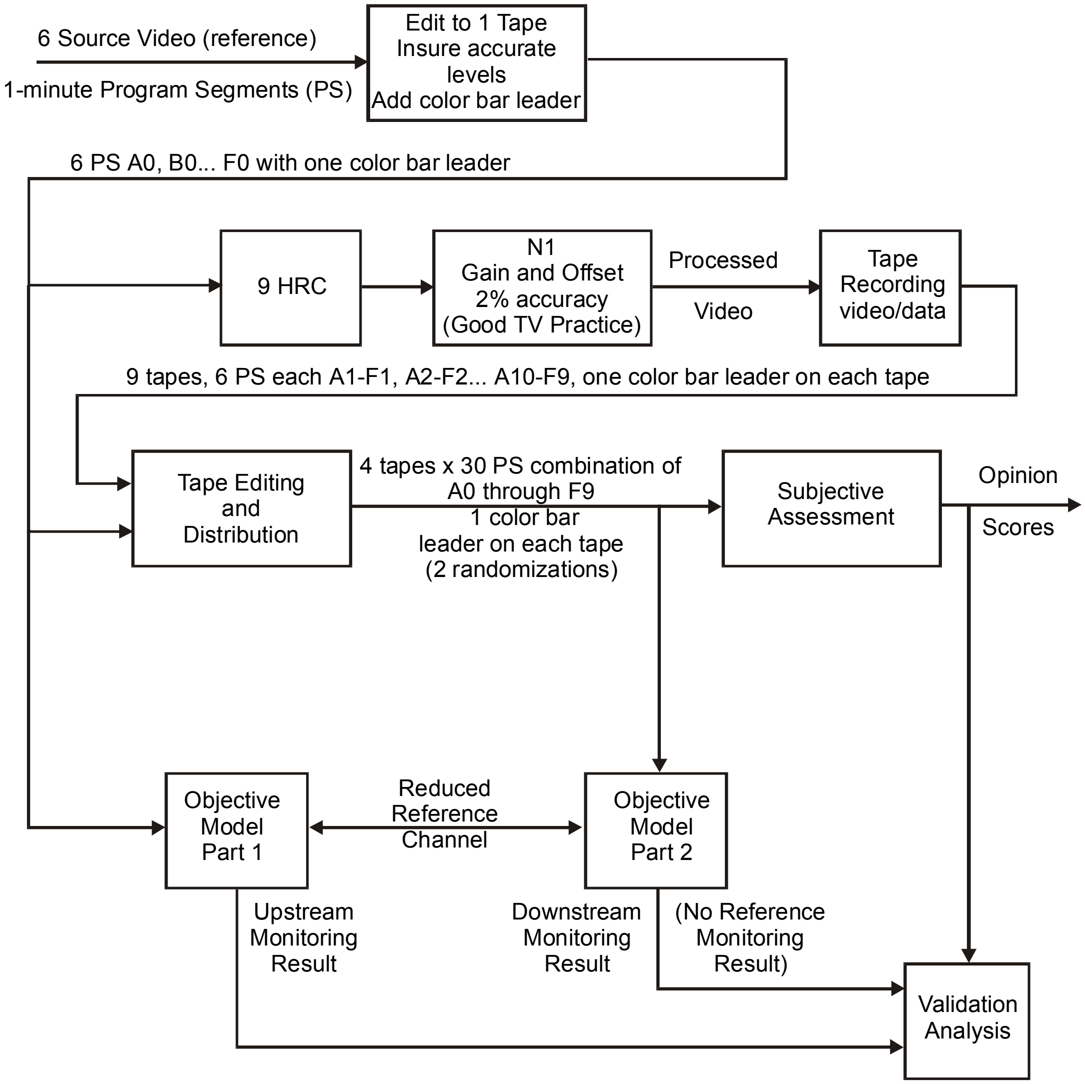
6 Source Video (reference)
1-minute Program Segments (PS)
Edit to 1 Tape
Insure accurate levels
Add color bar leader
6 PS A0, B0... F0 with one color bar leader
9 HRC
N1
Gain and Offset
2% accuracy
(Good TV Practice)
Processed
Video
Tape Recording video/data
9 tapes, 6 PS each A1-F1, A2-F2... A10-F9, one color bar leader on each tape
Tape Editing and Distribution
4 tapes x 30 PS combination of A0 through F9
1 color bar leader on each tape
(2 randomizations)
Subjective Assessment
Opinion
Scores
Objective Model Part 1
Reduced Reference Channel
Objective Model Part 2
Upstream Monitoring Result
Downstream Monitoring Result
(No Reference Monitoring Result)
Validation Analysis

Figure 3. Testing procedure overview.

1. 6 program segments (PS) are edited on to one tape. *Care*[3] is taken that amplitudes and levels are correct. One set of color bars is included as a leader to the tape. This produces PS A0 through F0.
2. Video from the A0-F0 tape is passed through 9 HRCs. *Care* is taken that amplitudes and levels are correct. One set of color bars is included as a leader to each tape. This produces 9 tapes with PS A1 through F1, A2 through F2, … , A9 through F9.

[3] *Care* means that the gain and offset of the video shall be adjusted to an accuracy of 2% of full amplitude either by adjustment within the HRC/VTR or by an external processing amplifier.

3. The 10 tapes are sources for production of the test session tapes (TS). This produces 4 tapes with 30 PS on each tape using A0 through F9 (two randomizations). One set of color bars is included as a leader to each tape for viewing monitor setup.
4. The A0 through F0 tape is used for production of the reference tape (RT, system source).
5. See section 4.1 for details on how these tapes will be used by the models.
6. PSNR will be calculated and reported if someone volunteers to do the calculation.

3.2. Test materials

3.2.1. Selection of test material

Six 1-minute segments (six for 525-line and six for 625-line) shall be selected taking into account the following considerations:

1. Preferably, each 1-minute scene should not have scene cuts more frequently than once every 10 seconds.
2. Objectionable material such as material with sexual overtones, violence and racial or ethnic stereotypes shall not be included. (N.B. cite text from Rec. 500)
3. The 1-minute scenes should each exhibit some range of coding complexity (i.e., spatial and temporal) within the 1-minute interval.
4. At least one scene must fully stress some of the HRCs in the test.
5. A portion of one scene should have some low level noise (e.g., ½- inch professional analog source).
6. No more than two of the six 1-minute scenes shall be from film source or contain film conversions.
7. No more than 30 seconds of one film scene shall contain 12 frames per second cartoon material.
8. The six scenes taken together should span the entire range of coding complexity (i.e., spatial and temporal) and content typically found in television.
9. If possible, one new SRC for each frame rate, which has not been disclosed to any proponent, each with its own set of 9 HRCs will be included by the ILG. If not possible, this requirement will be waived.

Video material currently available for the test:

Segment Gender	Characteristics	Currently Available Source
1. Sports	Fast motion	Men’s and Ladies’ Soccer, Volleyball, Dancing, Ballet
2. Winter Sports	High contrast	Universal Theme Park, “The Thing”
3. News Speaker	No motion	
4. B-grade Movie	Various Motion	“Frankenstein”
5. Commercial Break	High Speed Motion	Universal Theme Park
6. Movie-Special Effects	Synthetic pictures	“Apollo 13,” “Fast and Furious,” “Mummy Returns”
7. Cartoon	Synthetic pictures	“Woody Woodpecker,” “Casper,” “Land Before Time”
8. TV report	Low motion / Natural scenes	“Sahara,” New York
9. TV Shopping	Low motion	

Detailed description of available video material:

Available Source	Content Description	Original Format / Content Provider	Duration	480i60	576i
“Apollo 13”	Lift off scene: synthetic picture, fine detail, jerky motion	Original Film, telecined to 480i60 Universal Studios; POC: Teranex	00:03:12	X-D5	
Ballet Dancing	Indoor Ballet Dancing Couple, fast rapid movement	Original Film, telecined to 480i60 Kodak; POC: Teranex	00:01:54	X-D5	
“Casper”	Synthetic picture-digital CGI	12 fps original converted to film at 24 fps, telecined to 480i60 and 576i50 Universal Studios; POC: Teranex	00:03:58	X-D5	X-D
Dancing	Ballet Dancing	Captured in D5 German Broadcaster SWR/ARD; POC Teranex			X-D
“Frankenstein"”	Black and white original, “Bringing to life” scene	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:04:05	X-D5	X-D
Ladies Soccer	Fast motion, complete game, pans across crowds	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	≈ 02:04:00		X-D
“Land Before Time”	Synthetic picture	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:03:40	X-D5	X-D
“Live on the Edge”	Movie Trailer-Car chasing scene	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:01:54	X-D5	
Men’s Soccer	Fast motion, complete game, pans across crowds	Captured in D5 German Broadcaster SWR/ARD; POC Teranex	≈ 02:04:00		X-D
Movie	Crime Movie showing a pursuit scene	Original Film (16:9), telecined to 576i50 German Broadcaster; POC Teranex			X-D
“Mummy Returns”	Movie Trailer-special effects	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:01:51	X-D5	
New York	Views from a boat trip	Original Film (16:9), telecined to 576i50 German Broadcaster; POC Teranex			X-D
“Sahara”	Natural scenery, bugs, reptiles, sand storm, waterfall, nocturnal animals, fine detail	Original Film/HiDef—HD Down (3/2) insertion Mandalay Media Arts; POC: Teranex	01:54:00	X-D5	X-D
“The Thing”	Remake of original, Snow scenes, various Motion	Original Film, telecined to 480i60 and 576i60 Universal Studios, POC: Teranex	00:03:39	X-D5	X-D
Universal Theme Park	Varying motion, high contrast, full sunlight, water rides, inside rides, roller coaster	Capture with DigiBetaCam Teranex; POC: Teranex	00:24:46	X-D5	X-D
Volleyball	Indoor volleyball match	Captured in D5 German Broadcaster SWR/ARD; POC Teranex			X-D
“Woody Woodpecker”	Synthetic picture-traditional animation	12 fps original converted to film at 24 fps, telecined to 480i60 and 576i50 Universal Studios; POC: Teranex	00:03:49	X-D5	X-D

3.2.2. Hypothetical reference circuits (HRC)

The Hypothetical Reference Circuits are chosen to be representative of the most common practices in the field of digital TV broadcast networks, for each of 50 or 60 Hz frame rates. Two stages are taken into account:

- The MPEG encoding of original video, and multiplexing,
- The modulation stage for transmission purposes.

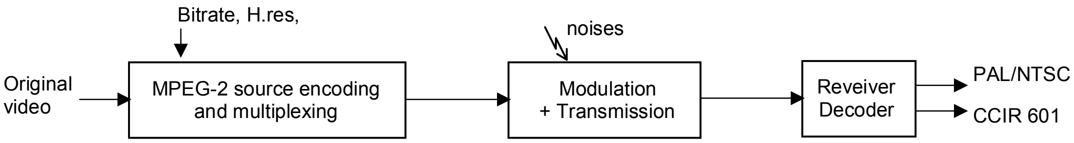

Figure 4. HRC generation chain

(N.B. Figure 4 should be revised after HRC selection.)

Although this chain appears simple, many configurations are possible. In order to limit the number of HRCs and the overall number of tests to be performed to a practical level, all combinations cannot be tested. Furthermore, the goal of these tests is to discriminate between the proposed models, not to study the impact of specific configurations on the perceived quality. As a consequence, the following directions should be adhered to:

1. Original digital signals are to be used.
2. At the encoding stage, a single encoding method should be chosen. The proposed range of encoding bit rates is 1 – 6 Mb/s plus variable bit rate achieved with statistical multiplexing.
3. At the transmission stage, many configurations and noises are possible.
    - Considering that all noises produce bit errors of varying lengths, only one modulation scheme should be retained. The 64-QAM (DVB-C) is a good candidate because the noise range from an error free output to no output at all at the receiver-decoder is wider than with other modulations (QPSK for example).
    - Several types of bit errors can be produced using two cases of white noise and one case of impulse noise.
4. At the receiving and decoding (IRD) stage, only one decoder should be selected. However, the output signal can be either digital (4:2:2 CCIR 601) or analogue (PAL/NTSC).

The proposed list of HRCs for preliminary coding is given in the table below. Nine HRCs will be selected from the 40 possible HRCs at a pre-selection meeting to be held at a later date. This pre-selection meeting will seek to obtain the widest distribution of quality. If the intended effect on quality is not achieved by any given HRC in the table, that HRC will be eliminated from consideration.

	Encoder [4]		Transmission		Decoder [5]	
HRC No.	Bit rates [Mbit/s]	H Res.	Mod.	Distortion	Output	Comments
0	-	Full	-	-	601	Original test signals
1-5	2, 4, 6 [4 & 2]	704	64-QAM	Q.E.F [8]	601	Several bit rates, QEF [6 rejected, 2 and 3 done]
6-10	2, 4, 6, [6 & 2]	528	64-QAM	Q.E.F	601	Pre-filtering influence
11–15	2, 4, 6, [6 & 2]	704	64-QAM	Q.E.F	PAL / NTSC	Analog output influence
16-20	2, 4, 6, [6 & 2]	704	64-QAM	WB noise (level 1)	601	Low level of transmission impairment
21-25	2, 4, 6, [6 & 2]	704	64-QAM	WB noise (level 2)	601	High level of transmission impairment
26-30	2, 4, 6, [6 & 2]	704	64-QAM	Impulsive noise	601	Impulsive transmission impairment
31-35	2, 4, 6, [6 & 2]	528	64-QAM	WB noise (level 1)	601	Low level of transmission impairment
36-40	2, 4, 6, [6 & 2]	704	64-QAM	Q.E.F	601	Several bit rates, QEF. Same as HRCs 1-5 except post-processing is included.

N.B. The table of 40 candidate HRCs will be modified to incorporate the following comments from the May 2001 VQEG meeting:

1. QAM is not required for Q.E.F.
2. May want to consider other error introducing mechanisms (e.g., ATM, etc.).

Ask experts to estimate quality levels of final selected set of HRCs (from the HRC selection meeting).

3. One high bit-rate HRC will have composite analog video processing (upstream processing preferred).
4. Some HRCs must be at 1 Mbits/s (poor quality).
5. At least one HRC should have pre-filtering.
6. Add a note that 704 is a truncation and 528 is a pre-filtering.

In order to generate the HRCs efficiently, the multiplexing capabilities of MPEG can be used. To that end, the HRCs that do not include transmission impairments can be multiplexed into a single Transport Stream. Then, the selected transmission configurations will be applied to this transport stream. In this

[4] As many (mp@ml) encoders as possible should be used but the encoder should be fixed for each HRC.

[5] As many decoders as possible should be used but the decoder should be fixed for each HRC.

[6] Variable (statistical multiplexing) where video bit rates should vary between 0.5 Mb/s to 8 Mb/s.

[7] Cascading of two MPEG coders.

[8] Quasi-Error Free.

way, the HRCs will be applied to all the original material (30 min) in 6+4 operations instead of 6×4. Each processed video segment will be identified in the following way: [seg. No “s”].[HRC No. “x.y”]. The 6×4 signals from the HRCs will be decoded and recorded on D-1 tapes.

Since the SSCQE evaluation method requires long duration sequences, assessing the whole material (6×4 sequences) for all HRCs could be very long. Thus, the test sequence will contain several HRCs applied only to a subset of the original sequence. The minimal length of an HRC is the segment length. The test sequences will be produced by video editing.

3.2.3. Segmentation of test material

The test video sequences will be in ITU Recommendation 601-2 4:2:2 component video format as described in SMPTE 125M, and recorded on D1 tapes. This may be in either 525/60 or 625/50 line formats. The temporal ordering of fields F1 and F2 will be described below with the field containing line 1 of (stored) video referred to as the Top-Field.

Video Data storage:

A LINE: of video consists of 1440 8-bit (byte) data fields in multiplexed order: Cb Y Cr [Y]. Hence there are 720 Y, 360 Cb and 360 Cr bytes per line of video.

A FRAME: of video consists of 486 active lines for 525/60 Hz material and 576 active lines for 625/50 Hz material. Each frame consists of two interlaced Fields, F1 and F2. The temporal ordering of F1 and F2 can be easily confused due to cropping and so it is constrained as follows:

> For 525/60 material: F1--the Top-Field-- (containing line 1 of FILE storage) is temporally LATER (than field F2). F1 and F2 are stored interlaced.
>
> For 625/50 material: F1--the Top-Field-- is temporally EARLIER than F2.
>
> The Frame SIZE:
> for 525/60 is: 699840 bytes/frame,
> for 625/50 is: 829440 bytes/frame.

A FILE: is a contiguous byte stream composed of sequences of frames as described above. For example, a 10 second length video sequence will have total byte counts:

> for 525/60 : 300 frames = 209,952,000 bytes/sequence,
> for 625/50 : 250 frames = 207,360,000 bytes/sequence.

Multiplex structure: Cb Y Cr [Y] ... 1440 bytes/line
720 Y/line
360 Cb/line
360 Cr/line

Format summary:

	-- 525/60 --	-- 625/50 --
active lines	486	576
frame size (bytes)	699840	829440
fields/sec (Hz)	60	50
Top-Field (F1)	LATER	EARLIER

3.2.4. Distribution of tests over facilities

Each test tape will be assigned a number so tracking of which facility conducts which test may be facilitated. The tape number will be inserted directly into the data file so that the data is linked to one test tape.

3.2.5. Processing and editing sequences

The video sequences will be Rec. 601 digital video sequences in either 625/50 or 525/60 format**. The choice of HRC's and Processing by the ILG will verify that the following operations do not occur between Source and Processed sequence pairs (excluding the non-compliant HRCs)**:

- Picture cropping greater than 10 pixels per side (excluding the 704 truncation)
- Chroma/luma differential timing
- Picture jitter
- Spatial scaling (size change)
- Horizontal shift greater than 5 pixels. (This criterion will be readdressed if such codecs are not available.)
- Vertical shift greater than 2 frame lines. (This criterion will be readdressed if such codecs are not available.)

Figure 5. is provided as an example HRC.A Rec. 601 Source component is passed through an MPEG-2 encoder at the various HRCs with the processed sequences recorded on a D1/D5 VTR.

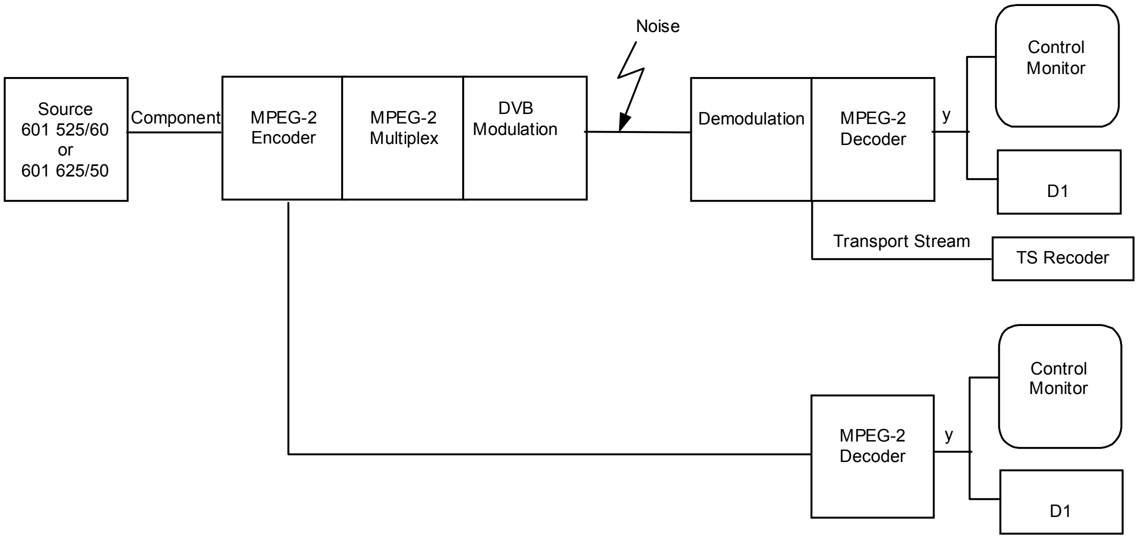

Figure 5. Example HRC

The source sequence is the MPEG-2 decoded sequence edited on D1 test tapes. The processed sequences are then edited onto D1 test tapes using edit decision lists leading to the repartition of impairments, distributed to each test facility for use in subjective testing sessions.

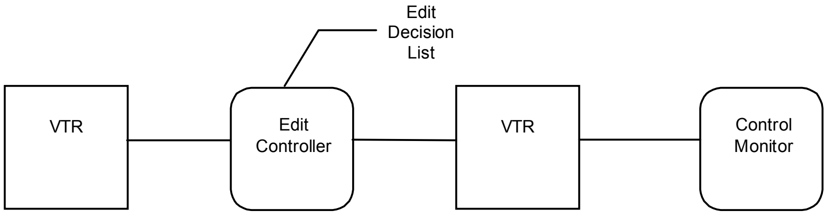

Figure 6. Edit processing

3.2.6. Randomization

For all test tapes produced, a detailed Edit Decision List will be created with an effort to:

- spread conditions and sequences evenly over tapes for any given session
- try to have a minimum of 2 trials between the same sequence
- have a maximum of 2 consecutive conditions, i.e. HRC's
- ensure that no sequence is preceded or followed by any other specific sequence more than once in order to minimise contextual effects

3.2.7. Presentation structure of test material

Due to fatigue issues, the session is limited to a 30 minute viewing period. For sessions conducted consecutively, there should be a minimum of a 15 minute break between sessions. It is recommended that both sessions be conducted on the same day. This will allow for maximum exposure and best use of any one viewer.

A training process will be used to instruct subjects about the task they are to perform during the subjective test and to bound the quality range to be seen in the test. An initial training process will be carried out before the beginning of the formal subjective test consisting of the dictation of instructions, the conduction of a training test session and a short time dedicated to question and answers (if any). The training test session will be repeated at the beginning of the second viewing session to reset the bounds of the quality range for the subjects.

3.3. Synchronization

3.3.1. Synchronization of data sampling with timecode

All subjective and objective data will be synchronized to the same timecodes for the duration of the test. Data will be produced at a rate of 2 samples per second.

3.3.2. Synchronization of source and processed sequences

It is important that synchronization be maintained between the one minute SRC and HRC sequences. Losses in synchronization may be the result of HRC processing delays, or the editing process itself.

To assure frame accurate synchronization, the SRC and HRC sequences will be visually matched at positions *first_frame* and *first_frame+n*, where *first_frame+n* is any suitable later transitional frame (scene cut) containing relatively high motion. The use of a high motion transitional frame allows the detection of even/odd field order inconsistencies, which can also be caused by HRC processing or videotape editing. It may be possible to correct these field order inconsistencies by forcing edits to occur on specific fields. The SRC and HRC *last_frame* positions should also be compared.

The SRC and HRC sequences shall be synchronized to within plus / minus 1 field. Subjective test tapes, and proponent video files, shall be derived from these matched SRC and HRC sequences.

4.Testing procedure

4.1. Model input and output data format

Upstream Model Original Video Side:

> The software for the original video side will be given the original test tape in the final file format to be used in the test, and a reference data file that contains the reduced-reference information (see Model Part 1 Processed Video Side).
>
> The software will produce an ASCII file, listing the Time Code of the original sequence, and the resulting video quality metric (VQM) of the model, with a resolution of 2 samples per second.

Upstream Model Processed Video Side:

> The software for the processed video side will be given the processed test tape in the final file format to be used in the test, and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bitrate of the reference data and consequently the class of the method (0, 10, 56 or 256 Kbits/s).

Downstream Model Original Video Side:

> The software for the original video side will be given the original test tape in the final file format to be used in the test, and produce a reference data file. The amount of reference information in this data file will be evaluated in order to estimate the bitrate of the reference data and consequently the class of the method (0, 10, 56 or 256 Kbits/s).

Downstream Model Processed Video Side:

> The software for the processed video side will be given the processed test tape in the final file format to be used in the test, and a reference data file that contains the reduced-reference information (see Model Original Video Side).
>
> The software will produce an ASCII file, listing the Time Code of the processed sequence, and the resulting video quality metric (VQM) of the model, with a resolution of 2 samples per second.

Note that all 4 video inputs/outputs need the information discussed in sections 3.3.1 and 3.3.2.

The first line of the ASCII file should contain the following string: <processed filename>

Each line of the ASCII file has the following format:

TimeCode VQM MOV_1 MOV_2 …. MOV_N

Where <processed filename> is the name of the processed sequence run through this model, without any path information, and VQM is the video quality estimation produced by the objective model. Each proponent is also allowed to add Model Output Values (MOV) that the proponent considers to be important. Only results of VQM calculations will be evaluated by comparative analysis.

4.2. Submission of executable model

The objective model should be capable of receiving as input the source sequence described in part 1, and the processed sequence corresponding to part 2, with the reduced reference data file. Based on this information, it must provide one unique figure of merit that estimates the subjective assessment value (VQM) of the processed material.

The objective model must be effective in evaluating the performance of block-based coding schemes (such as MPEG-2) in a range of bitrates between 1 Mb/s and 6 Mb/s on sequences with differing amounts of spatial and temporal information.

Proponents may submit up to 4 models, one for each of the reduced reference information bit rates given in the test plan (i.e., 0, 10 kb/sec, 56 kb/sec, 256 kb/sec).

The submission(s) should include a written description of the model including fundamental principles and available test results in a fashion that does not violate the intellectual property rights of the proponent. In order to be coherent with ITU work, the proponent model must be described in a manner such as that specified by ITU-R Rep. BT.2020-1.

The test tapes will be available in the final file format to be used in the test. MOS data for these tapes will be made available to proponents as soon as possible.

Each proponent will submit an executable of the model(s) and the results for a common piece of video material to the Independent Labs Group (ILG).. alternately, proponents may supply object code working on any of the computers of the independent lab(s) or on a machine supplied by the proponent. The ILG verifies the output of the model on this piece of video material prior to the running of the test. If there is a discrepancy, the proponent and ILG will work together to resolve the discrepancy.

IMPORTANT: tapes will be sent to proponents when the ILG is given ALL proponent's models. No model will be accepted after tape distribution.

5. Objective quality model evaluation criteria

5.1. Post-processing of data

5.1.1. SSCQE Subjective Data

Objective models will be compared against these three sets of subjective data:

- Raw SSCQE data set.
- Normalized SSCQE data set according to zero mean and unit variance per individual subjects
- Normalized SSCQE data set according to zero mean and unit variance, per individual subjects followed by computing $\Delta x = S_x\text{-}P_x$ for each processed sequence x, where P_x is the trace of the processed clip and S_x is the trace of the corresponding hidden reference clip. Processing of the one-minute clips in this manner will aid in the removal of contextual effects and compensate for the possibility that the original sequences might contain impairments (i.e. encoding artifacts or compression in the source).

5.1.2. Time alignment of subjective and objective data

The latency that results from viewer reaction times and slider "stiffness" is uninteresting and will not be evaluated by VQEG. After comparing subjective and objective data each model developer will be allowed to provide to the ILG one global time shift per viewing tape of their objective model time history data (i.e., VQM) with respect to the average mean opinion score (MOS) data from the subjective test.

5.1.3. Discarding first 6 seconds of each one-minute clip

Each one-minute clip on the viewing tape can come from HRCs with vastly different qualities. Discarding the first six seconds of each transition provides a period of time for the average viewer response data to stabilize. Thus, after the objective model data has been globally time shifted for each viewing tape (section 5.1.2), the first six seconds of each one-minute clip will be discarded and not considered for further analysis.

5.1.4. Amplitude scaling of objective data

Model evaluation will be done with and without linear amplitude scaling of the objective data.
Each objective model will be allowed one linear amplitude scaling function with respect to the subjective data. This amplitude scaling function must be monotonic over the full range of observed data (VQM, MOS) and must take one of the following three forms (the A's are free fitting parameters in the equations below):

1. Linear Polynomial: $MOS_p(VQM) = A0 + A1*(VQM)$.

Any use of amplitude scaling will be noted in the final report.

5.2. Introduction to evaluation metrics

A number of attributes characterize the performance of an objective video quality model as an estimator of video picture quality in a variety of applications. These attributes are listed in the following sections as:

- Prediction Accuracy
- Prediction Monotonicity
- Prediction Consistency

This section lists a set of metrics to measure these attributes. The metrics are derived from the objective model outputs and the results from viewer subjective rating of the test sequences. Both objective and subjective tests will provide a single number (figure of merit) for each half second of the processed sequence that correlates with the video quality MOS of the processed sequence. It is presumed that the subjective results include mean ratings and error estimates that take into account differences within the viewer population and differences between multiple subjective testing labs.

Evaluation metrics are described below and several metrics are computed to develop a set of comparison criteria. Furthermore, the data set should not be shared to keep information secure. Thus, if a proponent wanted to share the data set to distinguish several reduced reference bitrate categories, or other specific aspects, it will have to be discussed before the data analysis starts. **The data set parts will have to be large enough to allow relevant statistical analysis (at least 600 MOS corresponding to one segment). Finally, all data parts will be size equivalent, and have the same standard deviation, to be compared each other.**

Summary of evaluation criteria:

Metric 1	95% confidence interval
Metric 2	Root mean square error
Metric 3	Pearson linear correlation
Metric 4	Spearman rank order correlation
Metric 5	Outlier ratio
Metric 6	Kurtosis
Metric 7	Kappa coefficient
Metric 8	Resolving power
Metric 9	Classification errors

5.3. Evaluation Metrics

This section lists the evaluation metrics to be calculated on the subjective and objective data. The objective model prediction performance is evaluated by computing various metrics on the actual sets of data.

The set of differences between measured and predicted MOS is defined as the quality-error set Qerror[]:

$$\text{Qerror}[i] = \text{MOS}[i] - \text{MOS}_p[i]$$

Where the index *i* refers to a Time Code of the processed video sequence.

5.3.1. Metrics relating to Prediction Accuracy of a model

Metric 1: The **95% inverse-confidence interval-weighted root-mean-square error** of the error set Qerror[].

$$\sqrt{\left(\frac{1}{N}\sum_{N}\left(\frac{\text{Qerror}[i]}{\text{CONF}[i]+0,5}\right)^2\right)}$$

with CONF[i] = 95% confidence interval for the i^{th} point (of *N* points). The constant factor of 0,5 is added to stabilize the calculation for cases of very small confidence interval.

Metric 2: The simple **root-mean-square error** of the error set Qerror[].

$$\sqrt{\left(\frac{1}{N}\sum_{N}\text{Qerror}[i]^2\right)}$$

5.3.2. Metrics relating to Prediction Monotonicity of a model

Metric 3: **Pearson's correlation coefficient** between MOS and MOSp**.**
Metric 4: **Spearman rank order correlation** coefficient between MOS_p and MOS.

5.3.3. Metrics relating to Prediction Consistency of a model

Metric 5: **Outlier Ratio** of "outlier-points" to total points N.

Outlier Ratio = (total number of outliers)/N

where an outlier is a point for which: ABS[Qerror[i]] > 2*MOSStandardError[i].
Twice the MOS Standard Error is used as the threshold for defining an outlier point.

Metric 6:
Percentage of points outside the 95%CI. [redundancy check for this metric pending]

5.3.4. Metrics relating to agreement

Metric 7: The **Kappa** coefficient.

$$K=\frac{\sum_{i=1}^{m} f_o - \sum_{i=1}^{m} f_E}{N-\sum_{i=1}^{m} f_E}$$

Where f_{o} is the observed number of agreement between MOS and MOSp for each of the *m* MOS classes, and f_{E} is the number of agreement due to coincidence (can be computed by the product of the number of MOS and MOSp for a given class, divided by N).

	MOS 1	MOS 2	MOS 3	MOS 4	…	MOS *m*	Total
MOSp 1	fo(1)						Tp 1
MOSp 2		fo(2)					Tp 2
MOSp 3			fo(3)				Tp 3
MOSp 4				fo(4)			Tp 4
…					…		
MOSp *m*						fo(*m*)	Tp *m*
Total	T 1	T 2	T 3	T 4		T *m*	N

Where Ti = #MOSi and Tpi = #MOSpi,
m=100,
and, f_{E}(i)= (Ti×Tpi)/N.

So, the Kappa is a metric of agreement, and is not influenced by coincidence. Thus, K values are between –1 and 1, but do not have to be interpreted as a correlation coefficient. K values are lower than correlation, and a value around 0.4 indicates that the method is efficient.

5.3.5. Resolving Power and Classification Errors Evaluation Metrics

These methods are described in T1.TR.PP.72-2001 ("Methodological framework for specifying accuracy and cross calibration of video quality metrics") and will be computed , if possible, as a pilot auxiliary study.

5.4. Complexity

The performance of a model as measured by the above Metrics 1 – 7 will be used as the primary basis for model analysis. The specification of model complexity, while potentially important, is not in the scope of this test. This information can be requested from the proponents.

5.5. Objective results verification

The following procedure will be used to verify the results of the objective models before preparation of the final report.

1	Each proponent receives processed video sequences. Each proponent analyzes all the video sequences and sends the results to the Independent Labs Group (ILG).
2	The independent lab(s) must have running in their lab the software provided by the proponents, see section 4.2. To reduce the workload on the independent lab(s), the independent lab(s) will verify a random sequence subset (about 10%) of all video sequences to verify that the software produces the same results as the proponents within an acceptable error of 2%. The random subset will be selected by the ILG and kept confidential.
3	If errors greater than 2% are found, then the independent lab and proponent lab will work together to analyze intermediate results and attempt to discover sources of errors. If processing and handling errors are ruled out, then the ILG will review the final and intermediate results and recommend further action.
4	The model output will be the MOSp data set calculated over the sequence. The MOSp values are expected to correlate with the Mean Opinion Scores (MOS) resulting from the VQEG's subjective testing experiment.

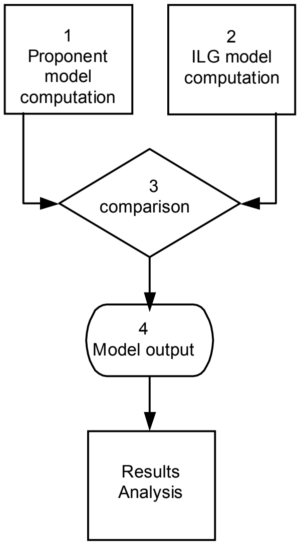

Figure 7. Results analysis overview.

6. Calendar and actions

Action	Due date	Source	Destination
Test plan final version	February 28, 2002	VQEG TDF-C2R	
Call for proposals	March 22, 2002	VQEG	Proponents
Sequence and HRC selection	May 15, 2002	ILG	
Fee payment	October 15, 2002		
Submission of executable models	October 15, 2002	Proponents	ILG
Sequence processing and tape editing	TBD	---	ILG
Video material delivered to proponents (?)	TBD		
Objective data delivered	TBD(4wks after receiving video material)	Proponents	ILG
Formal subjective test	TBD	ILG	
Subjective data analysis	TBD		
Objective data analysis	TBD		
Final report.	February 2003		

Note: Confirmation of model submission deadline will be made three weeks prior based on the availability of resources necessary to carry out the test.

7. Conclusions

VQEG will deliver a report containing the results of the objective video quality models based on the primary evaluation metrics defined in section 5. The Study Groups involved (ITU-T SG 9, and ITU-R SG 6) will make the final decision(s) on ITU Recommendations.

8. Bibliography

– VQEG Phase I final report.

– VQEG Phase I Objective Test Plan.

– VQEG Phase I Subjective Test Plan.

- VQEG FR-TV Phase II Test Plan.
- Recommendation ITU-R BT.500-10.
- ITU-R Report BT.2020-1.